"Smart Content Scraping" pour la construction de réseaux d'auteurs

Sonia Guérin-Hamdi
Institut des Sciences de l'Homme - ISH / CNRS
69363, Lyon, France
sonia.guerin-hamdi@ish-lyon.cnrs.fr

RESUME

« Mixed-Method Research On The Role of Research Communities in Environmental Politics » est un projet de recherche mené par une chercheuse en SHS de l'université de Tokyo (Japon), dont l'objectif est de mettre en évidence la relation de causalité entre les réseaux et les phénomènes sociaux (crise écologique, etc..).

Notre collaboration dans le cadre de ce projet vise avant tout à synthétiser le réseau d'auteurs hétérogènes à travers plusieurs types de relations, ceci pour une période donnée, incluant une crise écologique (« Dust Bowl », USA, the 1930's).

Dans cet objectif, nous nous intéresserons, d'une part, à un ensemble de données bibliographiques pour la construction d'un réseau de co-auteurs, et d'autre part, nous combinerons ces données avec le corpus textuel qui s'y réfère afin de considérer une autre dimension de ce réseau en mettant en évidence un autre type de relation, les relations de citation entre auteurs.

Nous nous appuierons sur les méthodes, outils et techniques de fouille de textes pour construire, fiabiliser, analyser et enrichir les réseaux.

Mots Clés :

Content scraping; semantic relations; citation index; text analysis; network analysis; named entity extraction;

ACM Classification Keywords

F.4.2 Grammars and Other Rewriting Systems: Parsing; H.3.1 Content Analysis and Indexing: Indexing methods; I.2.4 Knowledge Representation Formalisms and Methods: Semantic networks; I.2.7 Natural Language Processing: Text analysis; I.5.2 Design Methodology: Pattern analysis

INTRODUCTION

L'étude du phénomène social s'intéresse à la production scientifique et à l'activité de citation, au moyen de méthodes quantitatives, utilisant des outils statistiques et mathématiques, pour tenter de décrire, analyser, expliquer et prédire.

Le premier indice de citations(*Citation Index*) pour les articles publiés dans les revues scientifiques, a introduit, en 1960, par l'Institute for Scientific Information (ISI), et rendu accessible à travers le Web of Science, permettant ainsi d'établir facilement les liens entre documents cités et documents qui citent. Il s'agit d'abord du Science Citation Index (SCI), puis plus tard du Social Sciences Citation Index (SSCI) et de l'Arts and Humanities Citation Index (AHCI). Citons également, Google Scholar[1] ou CiteSeer^{x2} qui permet de rechercher par citation et d'ordonner les documents par l'impact des citations.

Une approche réseau met en évidence les relations existantes entre différentes entités, afin de structurer et analyser plusieurs réseaux tels que les réseaux de co-auteurs, et les réseaux de citation de publications. Un réseau d'entité est modélisé à l'aide d'un graphe composé d'un ensemble de nœuds et d'un ensemble de liens entre ces nœuds.

Dans notre étude, un nœud représente un auteur, et un lien exprime une relation de co-écriture ou de citation. Certains nœuds peuvent donc être reliés par une ou plusieurs relations. Nous distinguons les « citations explicites » extraites des références bibliographiques et les « citations implicites » extraites du corps des publications.

METHODOLOGIE

Les métadonnées des publications de notre corpus seront stockées dans une base de données relationnelle afin de faciliter leur extraction et construire un premier référentiel fiable d'auteurs.

Une bibliographie est constituée de références bibliographiques. Il s'agit d'une compilation de toutes les

[1] http://www.scholar.google.com/
[2] http://citeseerx.ist.psu.edu/

sources de données utilisées dans le cadre d'un travail de recherche. Chaque référence ou citation de publication scientifique renseigne sur des éléments précis (les métadonnées) caractérisant cette publication (auteur, titre, édition, année de publication, etc.). Ces métadonnées sont placées dans un ordre défini, caractérisées par une typographie particulière et séparées par une ponctuation normalisée. Les styles de citation diffèrent entre les disciplines, les éditeurs et les auteurs ce qui complique la tâche d'analyse de la citation.

Pour l'exploitation de la section « Bibiographie », nous avons retenu l'outil ParsCit[3]. Partant d'une publication en texte brut, ParsCit génère des données utiles de la publication, entre autres, la structure logique du document et les informations de la section « Bibliographie » liées à chaque référence bibliographique (par exemple : auteurs, titre, date). Il nous sera donc d'une aide précieuse pour extraire les métadonnées des références bibliographiques des papiers de notre corpus de texte. Ces informations seront en effet utilisées pour étoffer le référentiel d'auteurs précédemment construit, et structurer le réseau d'auteurs par les relations de citations « explicites » qui interconnectent autrement les auteurs identifiés.

A partir du référentiel construit, il sera possible de « fouiller » dans le contenu des papiers constituant le corpus pour détecter les citations implicites. En effet, le contenu des publications, extrait par ParsCit sera exploiter par le moteur d'indexation fulltext SolR, pour repérer les instances de chaque auteur et retourner le nombre d'occurrences d'instances trouvées.

Dans cet objectif, nous avons défini un processus de constitution de réseau de citations d'auteurs qui se décompose en 4 étapes :
1- Préparation des jeux de données, 2- réalisation d'un référentiel d'auteurs (identification des nœuds du graphe), 3- extraction de "citations explicites" et extraction de "citations implicites", développées dans la suite du document.

MISE EN ŒUVRE ET EXPERIMENTATION

Notre matériau de travail est décrit dans le Tableau 1. Il comprend :

- un corpus de papiers scientifiques correspondant à un échantillon de 449 publications issues de JSTOR dont la mission est de numériser des collections complètes de périodiques en SHS, économie et statistiques. Les papiers sont au format PDF, écrits en anglais et répertoriés par discipline : Ecologie, Anthropologie, Géographie et géologie. La fenêtre de publication se situe entre 1920 et 1950.
- un export des métadonnées (titre, date, revue de publication, auteurs, nombre de pages, url d'accès, JSTOR, ISSN, DOI.) pour le même échantillon de publications. sous forme d'un fichier sous format RDF fourni par Zotero (logiciel de gestion de références bibliographiques).

Disciplines	Nombre de publications
Ecologie	158
Géographie et Géologie	117
Anthropologie	34
Autres	140
TOTAL	**449**

Tableau 1. Le corpus extrait de JSTOR

Etape 1 : Préparation Des Jeux De Données

Sur un plan technique, des scripts de « Scraping », implémentés en PHP, nous ont permis d'extraire les métadonnées issues de l'export Zotero et d'alimenter notre base de données MySql. Nous avons ajouté les disciplines et le pointeur vers le texte intégral de chaque publication. Cette tâche de mise en correspondance consistait à rapprocher les enregistrements de la base de données et les noms de fichiers, se présentant comme une concaténation de métadonnées de la publication. Afin de s'affranchir d'un travail manuel fastidieux, nous avons utilisé « Soundex » un algorithme phonétique d'indexation de noms par leur prononciation en anglais britannique. L'idée étant de rapprocher les titres et les noms de fichiers correspondant, malgré des différences mineures d'écriture.

Le format PDF des publications n'est pas un format aisément exploitable à des fins d'extraction. Pour cette raison, nous avons transformé l'ensemble de ces fichiers dans un format plus adapté. Le format choisi est le texte brut, qui peut être utilisé comme format d'entrée pour la majorité des outils d'extraction. Nous avons implémenté des scripts Shell/PHP se greffant à ParsCit pour générer automatiquement, pour chaque publication, deux fichiers distincts, l'un se rapportant au corps (contenu) de la publication et l'autre lié à la bibliographie (références bibliographiques). Les analyses ultérieures porteront ainsi sur des jeux de données ciblés.

Un premier traitement s'est porté sur une vingtaine de publications, échantillon hétérogène représentatif de notre corpus initial, afin d'effectuer une série d'adaptation permettant d'améliorer la précision de l'outil ParsCit. Nous avons ensuite appliqué notre traitement sur l'ensemble du corpus de publications.

Etape 2 : Création du Premier Référentiel d'Auteurs

Une ébauche de référentiel d'auteurs a rapidement été constituée à partir de la liste des auteurs renseignés dans les métadonnées des publications référencées dans la base de données précédemment construites. Un auteur se caractérise selon plusieurs attributs : Nom / prénom(s) / discipline(s) / année(s) de publication

[3] http://aye.comp.nus.edu.sg/parsCit/

Etape 3 : Bibliographie et « Citations Explicites »
L'outil ParsCit nous fournit les fichiers regroupant les références bibliographiques des publications de notre corpus sous format XML, nous facilitant ainsi grandement la tâche d'analyse.

En effet, les informations sont structurées et il devient alors aisé de parcourir via un script (PHP) ces fichiers pour en extraire les informations concernant les auteurs cités et les stocker dans notre base de données relationnelle. Ce script a été appliqué sur l'ensemble des documents de notre corpus de travail. Le résultat permet, d'une part, d'enrichir le référentiel des auteurs des publications initialement créé (nous passons de 331 à 400 auteurs) et d'autre part, d'identifier les relations de « citations explicites » entre les auteurs.

SMART SCRAPING ET « CITATIONS IMPLICITES »

A cette étape du processus, il est question d'extraire l'ensemble des citations dites « citations implicites » identifiées dans le corps du texte.
Dans cette partie, nous nous intéressons au corps de la publication (la publication sans sa bibliographie). Il s'agit d'explorer les contenus pour extraire les auteurs cités directement dans le contenu. Nous avons utilisé le référentiel des auteurs précédemment enrichi incluant des auteurs cités explicitement dans les références bibliographiques.

Indexation par SolR, Moteur de Recherche Fulltext
Un moteur d'indexation, tel que SolR[4] sait reconnaitre un « pattern » dans le corps des publications et calculer le « Term Frequency », la fréquence du terme dans la publication.

Lors d'un premier test, nous avons recherché la présence des « fullnames » (Prénom Nom) issus du référentiel. Les résultats obtenus (373 liens) n'affichaient que peu de citations détectées. En effet, pour citer un auteur, il est parfois utilisé le nom et le prénom complet, parfois seulement l'initiale du prénom apparait, et d'autres fois aucun prénom n'est indiqué.

Dans un second temps, nous nous sommes intéressés aux « surnames » (noms de famille) uniquement. Les résultats (5332 liens) démontrent beaucoup de bruits et nous confrontent à d'autres difficultés que nous devrons explorer plus précisément. Nous avons pour le nom « Wright », 3 instances dans notre référentiel auteur : « Georges M. Wright », « Elnora A. Wright », et « John C. Wright ». De même pour le nom « Brown », pour lequel 3 instances sont identifiées : « H. Ray Brown », « Leo Brown », « Ralph H. Brown ».

Author Name Desambiguation
A la question, Comment distinguer les homonymes ? Nous proposons une méthode consistant en une tâche semi-automatique de redistribution des citations aux auteurs concernés. Nous interrogeons le moteur d'indexation SolR, pour les noms de famille qui présentent des homonymes. Le moteur nous retourne pour un nom de famille soumis, les articles concernés et pour chacun le nombre de fois où l'auteur est cité dans l'article. Chaque instance trouvée est surlignée et son enveloppe (préfixe et suffixe) dans la phrase est identifiée. Chaque citation est alors attribuée à l'auteur concerné.

A partir de là, nous pouvons observer un réseau plus représentatif de l'activité de citation des auteurs relevés dans notre corpus.

EVALUATION PRELIMINAIRE

Dans le but d'analyser le réseau d'auteurs, nous avons fait le choix de le représenter sous la forme d'un graphe, utilisant pour sa construction un algorithme ForceAtlas. Les nœuds se repoussent, tandis que les liens attirent les nœuds qu'ils connectent tels des ressorts. Cet algorithme fait le choix de garder proche d'un nœud, les nœuds qu'il connecte [1].

Construction du Graphe pour le Visualisation
Pour une représentation graphique interactive sur une interface Web, nous utilisons le Framework JavaScript « Vis.js » qui exploite aisément un objet JSON, qui propose une structuration sous forme de graphe.

Relation de co-écriture
Sur une fenêtre temporelle 1920–1950, nous représentons sur la Figure 1, les 331 auteurs des papiers connectés par les relations de co-écriture identifiées. Il est intéressant d'observer que seules 47 publications sur les 449 publications du corpus sont concernées par ce type de relation. Et nous pouvons noter que seules 81 relations de co-écriture ont pu être extraites du corpus. Le corpus est très peu marqué par ce type de relation. Comme le montre la Figure 1, les sous-réseaux constitués sont cloisonnés et ne communiquent pas.

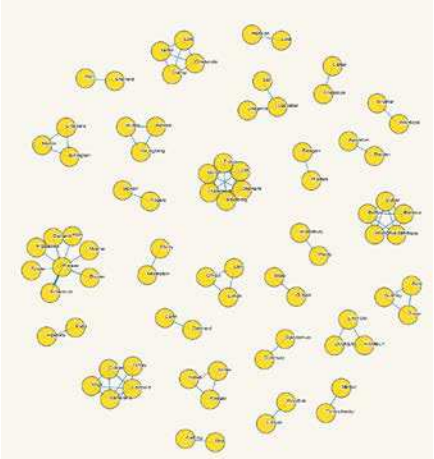

Figure 1. Réseau d'auteurs avec les relations de Co-écriture.

[4] http://Lucene.apache.org/solr/

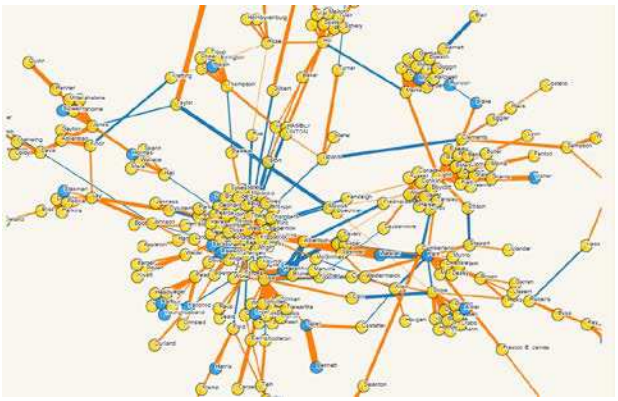

Figure 2. Réseau d'auteurs avec les relations de citations

Relation de citation

Le réseau d'auteurs interconnectés par les relations de citations se présente sous la forme d'une matrice asymétrique, traduite au format JSON pour sa représentation. Toujours sur une fenêtre temporelle 1920–1950, nous représentons sur la Figure 2, nos 400 auteurs reliés par des relations de citations. Un auteur très fréquemment cité, aura une position plutôt centrale, avec un grand nombre d'auteurs satellite. Nous avons fait le choix de Relations non dirigées, avec une relation pondérée sur le nombre de fois où l'auteur sera cité dans les articles du corpus par un autre auteur.

Cette représentation propose une catégorisation des nœuds par discipline d'appartenance, une catégorisation des relations par type de relation à dominante « implicite » ou « explicite ».

Une représentation web offre un rendu dynamique, donnant la possibilité de modifier les paramètres à la volée, permettant également une interactivité avec l'utilisateur qui explore le réseau avec sa souris. Par exemple, au survol d'un nœud, il peut obtenir les caractéristiques de l'auteur concerné, au clic sur un nœud, il peut observer ses voisins, auteurs avec lesquels il a construit une relation de citation. L'utilisateur peut également zoomer afin d'explorer plus précisément un sous-réseau ou tester la force de répulsion et d'attraction d'un nœud central.

Dans un souci de proposer une méthode complète et non biaisée, les citations implicites ne doivent pas être négligées car ces relations interconnectent fortement les nœuds du réseau. Elles s'avèrent même essentielles à identifier au vu de certaines pratiques dans certaines disciplines. En effet, si nous considérons la géographie et l'anthropologie, les références bibliographiques sont absentes.

La question de la représentativité du réseau d'auteurs que nous avons établi se pose. Nous pouvons faire l'hypothèse d'un recouvrement fort entre les activités scientifiques des disciplines de notre corpus, ou bien explorer les textes bruts des publications pour en extraire les entités nommées afin de s'assurer de ne rien manquer.

Named Entity Extraction APIs

Des extracteurs sémantiques proposant des APIs en ligne ont été exploités, en particulier pour la détection d'entités nommées dans le corps du texte. Pour la plupart, ces outils restent performants tant que le contenu reste dans la langue anglaise.

Notre attention s'est portée sur AlchemyAPI[5] et OpenCalais[6], tous deux accessibles en mode Saas (Service as a software). Chaque outil propose une API REST pour un grand nombre de fonctionnalités en text-mining et analyse de contenu. Ils possèdent des ontologies qui vont permettre entre autres d'identifier la langue du texte et les entités « Person » présentes dans le contenu du texte. D'autres informations portant sur les entités « City », « Country », « Organization »... sont fournies et pourront être utilisées pour d'autres études.

Pour tester les services, nous avons soumis des fichiers au format TXT (possibilité d'entrées HTML, XML, TXT, ou une URL) et récupérer les réponses sous forme XML ou JSON (sorties XML, JSON, RDF possibles). Le format JSON nous intéresse particulièrement, car il est aisément exploitable, représente un graphe, et non une arborescence (XML) avec l'avantage de fournir un support pour une écriture simple et légère au format texte, relativement compréhensible.

AlchemyAPI

Ce service fournit une extraction d'entités nommées et des capacités de désambiguisation pour l'analyse du texte. Dans l'ensemble, la qualité de l'entité est bonne, mais le nombre d'entités retournées est faible. Dans l'exemple (voir Code exemple 1.), les entités « Person» détectées extraites sont retournées avec le nombre d'occurrence des instances de l'entité détectée.

```
<entity>
   <type>Person</type>
   <relevance>0.250294</relevance>
   <count>2</count>
   <text>L. A. Stoddart</text>
</entity>
<entity>
   <type>Person</type>
   <relevance>0.237201</relevance>
   <count>1</count>
   <text>R. W. Darland</text>
</entity>
```

Code exemple 1. Exemple de résultat retourné par AlchemyAPI au format XML.

OpenCalais

Un produit de Thomson Reuters, fournit un moteur de traitement du langage naturel robuste pour extraire les

[5] http://www.alchemyapi.com/api/

[6] http://www.opencalais.com/documentation/calais-web-service-api/

entités sémantiques du texte, avec des capacités fortes de désambiguisation. Les entités fournies par le service Web de Calais sont pertinentes, et de bonne qualité. Mais le service ne supporte pas la saisie de documents texte supérieure à 100K caractères. Il est donc nécessaire de paramétrer le script d'interrogation pour tronquer les textes avant traitement par le service d'extraction.

Dans l'exemple (voir Code exemple 2), nous obtenons, pour un extrait de texte brut, une entité personne « R. L. Fowler » désambiguïsée. Le service OpenCalais, nous retourne également toutes les instances de cette entité détectée, ainsi que les parties de phrases l'enveloppant (préfixe et suffixe). Ces informations pourront être exploitées de manières plus approfondies pour consolider le référentiel d'auteurs caractérisant le corpus et fiabiliser les relations de citations implicites articulant le réseau. Notons toutefois que les extracteurs sémantiques restent performants lorsque le contenu textuel reste dans la langue anglaise.


```
"http://d.opencalais.com/pershasha1/xxxxx": {
   "_typeGroup": "entities",
   "_type": "Person",
   "name": "R. L. Fowler",
   "persontype": "N/A",
   "nationality": "N/A",
   "commonname": "R. L. Fowler",
   "_typeReference":
"http://s.opencalais.com/1/type/em/e/Person",
"instances": [
   {
   "detection":  "[by  Robertson  ('39)  and
Weaver,  Robertson,  and  ]Fowler[  ('40)  added
further  information.  Finally,  "pre□x":  "by
Robertson ('39) and Weaver, Robertson, and ",
   "exact": "Fowler",
   "su-x":  "  ('40)  added  further  information.
Finally, a",
   "offset": 3519,
   "length": 6
   },
],
"relevance": 0.577
}
```


Code exemle 2. Exemple de résultat retourné par OpenCalais au format JSON.

Nous obtenons pour un document texte soumis une entité personne « Fowler » désambiguïsée. Le service OpenCalais, nous retourne également toutes les versions de cette entité détectée, ainsi que les morceaux de phrases l'enveloppant (préfixe et suffixe). Ces informations pourront être exploitées de manières plus approfondies pour consolider le référentiel d'auteurs caractérisant le corpus et fiabiliser les relations de citations implicites articulant le réseau. Notons toutefois que les extracteurs sémantiques restent performants lorsque le contenu textuel reste dans la langue anglaise

Conclusion et poursuite des travaux

Nous pouvons maintenant définir un réseau d'auteurs selon 3 types de relations : relation de " co-écriture ", relation de "citations explicites" et relation de "citations implicites".

Le référentiel d'auteurs constitué et enrichi par les analyses et les extractions réalisées est essentiel. Il conviendra d'approfondir les méthodes et techniques à mettre en œuvre pour consolider ce référentiel d'auteurs.

Nous le complèterons par une détection automatique des entités nommées présentes dans les contenus des publications. A ce stade, il sera nécessaire de faire intervenir un expert du domaine pour valider ou rejeter les propositions provenant d'outils de détection automatique, afin de fiabiliser les résultats.

Pour la poursuite des analyses, nous pourrions exploiter, l'interaction entre les auteurs intra et inter discipline, ce qui permettrait de se rendre compte du recouvrement de l'activité scientifique entre les disciplines et comprendre comment elles interagissent entre elles.

Les premiers résultats nous montrent l'émergence de sous-réseaux. Nous pourrons effectuer une analyse diachronique de ces sous-réseaux, observer leur évolution dans le temps.

Ce type exploration devient possible grâce à l'utilisation d' « agrégats », appelés encore « facettes », (par année, par discipline, par requête personnalisée). Pour cela, nous nous intéresserons alors à un moteur d'indexation tel que SolR, qui permet cette navigation par facettes.

BIBLIOGRAPHIE

1. Jacomy M., Heymann S., Venturini T., Bastian M. ForceAtlas2, A Continuous Graph Layout algorithm for Handy Network Visualization, 2012.
2. Liu X., Bollen J., Nelson M.L., Van de Sompel H. Co-authorship networks in the digital library research community, 2005, 1462-1480.
3. Stern Rosa. Identification automatique d'entités pour l'enrichissement de contenus textuels. Thèse de Doctorat, France, 2014, tel.archives-ouvertes.fr/docs/00/93/94/20/PDF/these.pdf
4. Venturini Tommaso, Gemenne François, Severo Marla. Des migrants et des mots : Une analyse numérique des débats médiatiques sur les migrations et l'environnement, Cultures et Conflits n°88, 2012, 7-30.
5. Williams Kyle, Wu Jian, Choudhury Sagnik Ray, Khabsa Madian & Lee Giles C. Scholarly Big Data Information Extraction and Integration in the CiteSeer^{x} Digital Library. ICDE Workshop, 2014, 68-73.